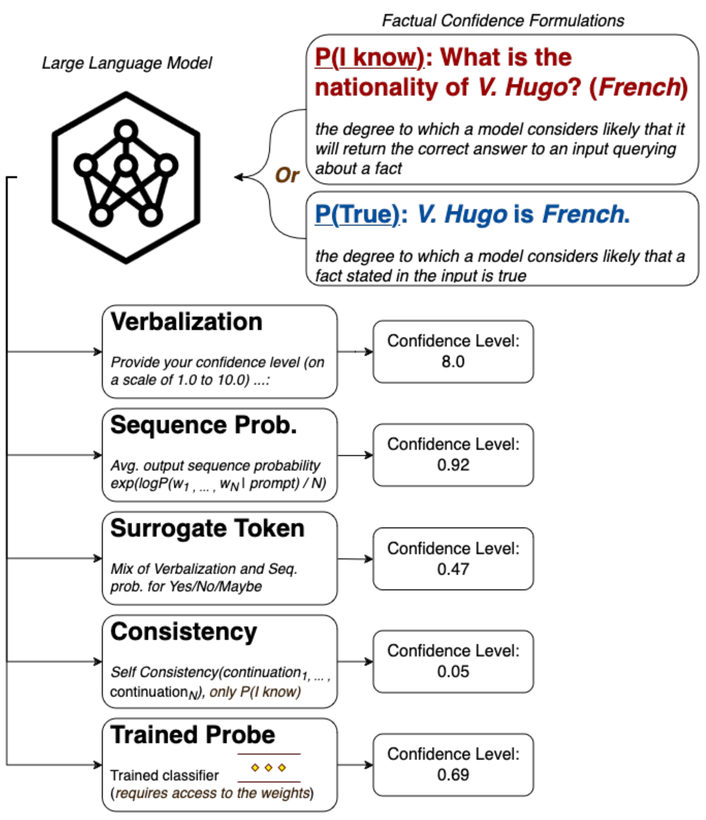
Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators

Matéo Mahaut[1*], Laura Aina[2], Paula Czarnowska[2], Momchil Hardalov[2], Thomas Müller[2], Lluís Màrquez[2]
mateo.mahaut@upf.edu, {eailaura, czarpaul, momchilh, thomul, lluismv}@amazon.com

1-Universitat Pompeu Fabra, 2-AWS AI Labs
*work done during an internship at AWS AI Labs

Problem definition and setup

Large Language Models (LLMs) tend to be unreliable in the factuality of their answers. Using atomic facts, we compare existing self-evaluation methods for factual confidence. We check resistance to meaning preserving perturbations (paraphrases /translation).



Datasets

Lama Trex: 34k triplets <subject, relation, object> from wikipedia *E.g.: Victor Hugo was born in France*. 34k false facts where objects are switched from the same relation E.*g.: Victor Hugo was born in Thaïland.*

PopQA: 14k questions + answers from wikipedia: *E.g. What is George Rankin's occupation? Answer: Politician.* Subjects and objects have a popularity metric: nº visits on wikipedia page of entity. Questions in PopQA are about low popularity entities.

80% of data is kept for training 20% is used at test time.

Models

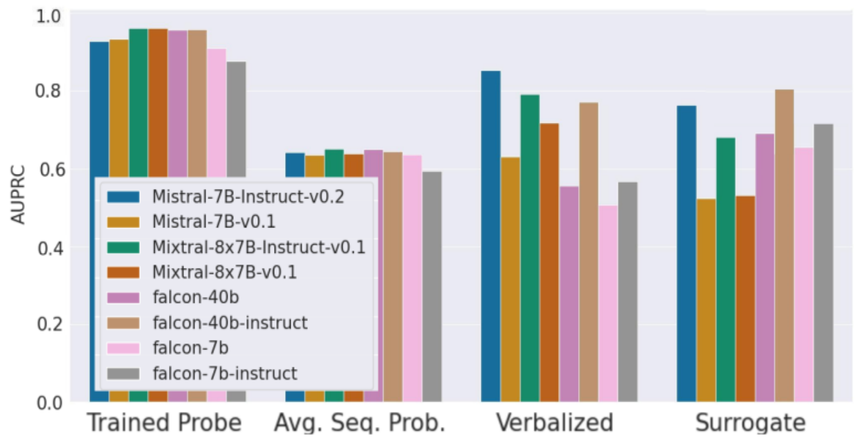
Names	Size	Open	Arch.	Instruct
Falcon	40B	✔	Dense	
Falcon Inst.	40B	✔	Dense	✔
Falcon	7B	✔	Dense	
Falcon Inst.	7B	✔	Dense	✔
Mixtral	46.7B	✔	SMoE	
Mixtral Inst.	46.7B	✔	SMoE	✔
Mistral	7B	✔	Dense	
Mistral Inst.	7B	✔	Dense	✔

Factual confidence estimators

	Black-box	Trained	Prompt-based	Scores for
Trained Probe	No	Yes	No	$P(\text{T})$ & $P(\text{IK})$
Sequence Probability	Yes (*)	No	No	$P(\text{T})$ & $P(\text{IK})$
Verbalization	Yes	No	Yes	$P(\text{T})$ & $P(\text{IK})$
Surrogate Token Probability	Yes (*)	No	Yes	$P(\text{T})$ & $P(\text{IK})$
Consistency	Yes	No	No	$P(\text{IK})$

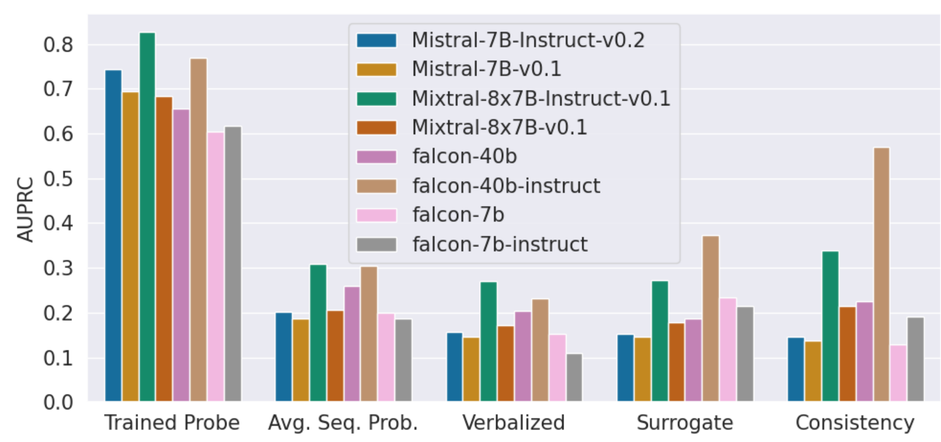
Results

AUPRC for P(T) on Lama Trex



All scoring methods perform above random. Trained probe and Average Sequence Probability have model-independent performance, while the prompt-based methods show a model-size and instruction-tuning effect.

AUPRC for P(IK) on PopQA



Trained Probe has a clear advantage on all other methods. Performance for other methods is barely above chance. Some outlier behaviour hints at impact of training data.

Generalization Analysis

P(T) Generalization to Unseen Dataset

Name	Size	AUPRC	Δ
Falcon	40B	.80	-.16
Falcon Ins.	40B	.81	-.15
Falcon	7B	.66	-.25
Falcon Ins	7B	.59	-.28
Mixtral	46.7B	.78	-.18
Mistral	7B	.62	-.31
Mistral Ins	7B	.75	-.18

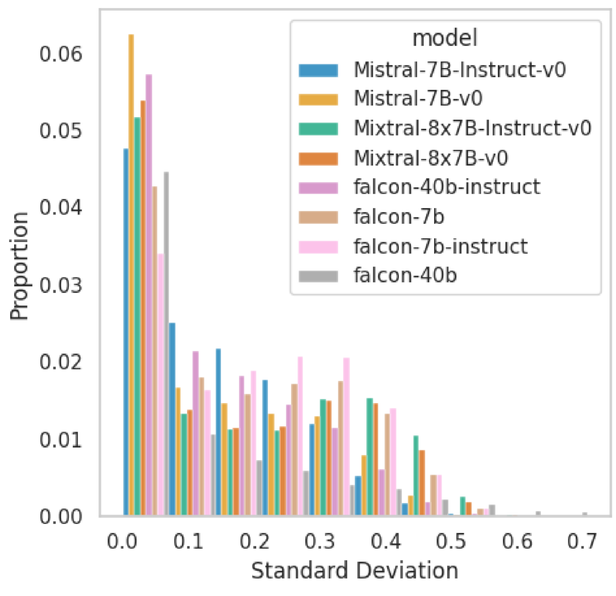
Out of domain AUPRC for trained probe method. Probe trained on Lama and tested on PopQA as true/false statements. Δ is difference with in-domain AUPRC .

P(T) Generalization to Translation

Name	Size	En-Fr	En-Po
Falcon	40B	.90	.86
Falcon Ins.	40B	.92	.87
Falcon	7B	.79	.44
Falcon Ins	7B	.67	.35
Mistral	7B	.67	.58
Mistral Ins	7B	.65	.53
Mixtral	46.7B	.87	.77

Pearson correlation of trained probe score for the same fact in different languages (English, French, and Polish). Zero-shot transfer to translated facts is within .10 of in domain AUPRC.

Variation of P(T) to Paraphrase



Variation of scores for artificially generated paraphrases of a given fact. Instabilities remain, which are only weakly explained by object popularity, and triplet relation. Factuality of an answer is not disentangled from words used to describe the fact.

Take-Home Message

We test 5 method types across model size, architectures and datasets, for P(T) and P(IK). Experiments show:

- Trained-probe based estimators perform better for both P(T) and P(IK) estimation. Its reliance on weights being accessible limit its potential use-cases.
- Trained probe generalizes well, but is not entirely robust to meaning preserving perturbations.
- All other methods have much lower performance, especially for non fine-tuned and smaller models. Prompt based methods are model dependant.

Future works are needed to strengthen those estimators, and improve robustness to meaning preserving perturbations.